FTA Report No. 0218

FTA RESEARCH
FEDERAL TRANSIT ADMINISTRATION

Emerging Data Science for Transit: Market Scan and Feasibility Analysis

PREPARED BY

David Perlman
Kristin Tufte
Lafcadio Flint
Tara Reel
John A. Volpe National Transportation Systems Center

U.S. Department of Transportation
Federal Transit Administration

JUNE
2022

COVER PHOTO

Courtesy of istockphoto.com

DISCLAIMER

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for its contents or use thereof. The United States Government does not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the objective of this report. The opinions and/or recommendations expressed herein do not necessarily reflect those of the U.S. Department of Transportation.

Emerging Data Science for Transit: Market Scan and Feasibility Analysis

JUNE 2022

FTA Report No. 0218

PREPARED BY

David Perlman
Kristin Tufte
Lafcadio Flint
Tara Reel
John A. Volpe National Transportation Systems Center
U.S. Department of Transportation
55 Broadway
Cambridge, MA 02142

SPONSORED BY

Federal Transit Administration
Office of Research, Demonstration and Innovation
U.S. Department of Transportation
1200 New Jersey Avenue, SE
Washington, DC 20590

AVAILABLE ONLINE

https://www.transit.dot.gov/about/research-innovation

Metric Conversion Table

SYMBOL	WHEN YOU KNOW	MULTIPLY BY	TO FIND	SYMBOL
LENGTH				
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
VOLUME				
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	cubic meters	m^3
yd^3	cubic yards	0.765	cubic meters	m^3
NOTE: volumes greater than 1000 L shall be shown in m^3				
MASS				
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams (or "metric ton")	Mg (or "t")
TEMPERATURE (exact degrees)				
^oF	Fahrenheit	5 (F-32)/9 or (F-32)/1.8	Celsius	^oC

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
June 2022	Final Report	September 2020–February 2022

4. TITLE AND SUBTITLE
Emerging Data Science for Transit: Market Scan and Feasibility Analysis

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
David Perlman, Kristin Tufte, Lafcadio Flint, Tara Reel

5d. PROGRAM NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESSE(ES)
John A. Volpe National Transportation Systems Center
U.S. Department of Transportation
55 Broadway
Cambridge, MA 02142

8. PERFORMING ORGANIZATION REPORT NUMBER
FTA Report No. 0218

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Department of Transportation
Federal Transit Administration
Office of Research, Demonstration and Innovation
1200 New Jersey Avenue, SE, Washington, DC 20590

10. SPONSOR/MONITOR'S ACRONYM(S)
FTA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12 . DISTRIBUTION/AVAILABILITY STATEMENT
Available from: National Technical Information Service (NTIS), Springfield, VA 22161; (703) 605-6000, Fax (703) 605-6900, email [orders@ntis.gov]; Distribution Code TRI-30

13. SUPPLEMENTARY NOTES [www.transit.dot.gov/research-innovation/fta-reports-and-publications] [https://www.transit.dot.gov/about/research- innovation] [https://doi.org/10.21949/1520704] Suggested citation: Federal TransitAdministration. Emerging Data Science for Transit: Market Scan and Feasibility Analysis. Washington, D.C.: United States Department of Transportation, 2021. https://doi.org/10.21949/1520704.

14. ABSTRACT

This report describes the state of the practice in the use of emerging data science tools and methodologies among U.S. transit agencies. It identifies commonalities in the tools and methods being used, as well as in the types of problems that agencies are seeking to solve using increasingly advanced data science approaches. The report also summarizes common challenges, opportunities, issues, and potential solutions among agencies that appear at the forefront in investing in emerging data science capabilities. Finally, the report outlines key factors and considerations for the further adoption of emerging data science practices and tools within the U.S. domestic transit industry.

15. SUBJECT TERMS
Data, data science, analysis, analytics, machine learning, innovation

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified	Unlimited	43	19b. TELEPHONE NUMBER

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

TABLE OF CONTENTS

LIST OF TABLES

Abstract

This report describes the state of the practice in the use of emerging data science tools and methodologies among U.S. transit agencies. It identifies commonalities in the tools and methods being used, as well as in the types of problems that agencies are seeking to solve using increasingly advanced data science approaches. The report also summarizes common challenges, opportunities, issues, and potential solutions among agencies that appear at the forefront in investing in emerging data science capabilities. Finally, the report outlines key factors and considerations for the further adoption of emerging data science practices and tools within the U.S. domestic transit industry.

Acknowledgments

This report was prepared by the Technology, Innovation, and Policy Division of the U.S. Department of Transportation (USDOT) John A. Volpe National Transportation Systems Center under sponsorship of the Federal Transit Administration (FTA) Office of Research, Demonstration and Innovation. The authors thank the project team from FTA, particularly David Schneider, Murat Omay, and Mohammed Yousuf. The authors also wish to thank the individuals interviewed for this report, who offered insights based on their experience in transit agencies across country, big and small.

Executive Summary

Public transportation providers in the United States collect a large volume of data through their daily operations, particularly through Intelligent Transportation Systems (ITS) technologies such as Automatic Vehicle Locators, Automatic Passenger Counters, and Automatic Fare Collection equipment. Transit agency staff are increasingly interested in leveraging these data to provide insights to enhance their understanding of community needs and improve planning, operations, and safety and are turning to data science for answers. However, data science is a relatively new field in public transportation, and agencies are still exploring its relevance and potential.

This report discusses practical opportunities for the use of data science in public transportation. The study identifies common tools and methods and the types of problems agencies are seeking to solve using increasingly advanced data science approaches. It also summarizes common challenges, opportunities, issues, and potential solutions among agencies that appear at the forefront in investing in emerging data science capabilities and outlines key factors and considerations for further adoption.

Through a literature review and interviews with transit agencies, the authors identified examples of advanced and emerging and data science, including applications to support:

- Asset health monitoring and predictive maintenance
- Occupant counting and monitoring
- Improving operational efficiency and information availability
- Planning, scheduling, and performance management

Many emerging applications focus more on data hygiene, management, and integration than on applying the latest cutting-edge analysis methodologies, although the agencies interviewed offered examples of targeted, problem-focused pilots and initial deployments of applications leveraging machine learning, computer vision, and other innovative tools.

A key finding of this study relates to critical factors in the adoption of expanded data analytics, including knowledgeable staff, effective data hygiene and data integration as a precursor to advanced analysis, and access to qualified vendors for assistance.

Section 1 Introduction

Motivation

Over the past several decades, public and private sector business have achieved greater access to data and information through data science technologies and methods. Companies in many sectors use data science to form theories, test hypotheses, find patterns, make predictions, and visualize new insights. Public transportation providers now generate broader and deeper data from their Intelligent Transportation Systems (ITS) technologies. Data generated by systems such as Automatic Vehicle Location (AVL), Automatic Passenger Counting (APC), Automated Fare Collection (AFC), and others can yield new information, enhance planning and management, and support data-driven decision making.[1] Agency staff can also use these datasets to perform exploratory analysis and obtain insights. Transit agencies have expressed interest in applications of open data and the potential benefits of data fusion and data mining, video information, freeform text analysis, and other emerging data sources and analysis methods. However, obtaining and managing data can present obstacles to taking full advantage of these new opportunities.

Prior research on this topic, including an ITS America report, noted that "the Transit Industry is not fully using the data it collects and is not yet positioned to expand its use in a future of ubiquitous data including Big Data, Smart Cities, and Connected Vehicles and Infrastructure" (1). As interest in data science and related fields—big data, artificial intelligence, and machine learning—grows, so has the hype surrounding them.

This report identifies and explores *practical* opportunities in these fields for public transportation while acknowledging where expectations may be inflated or beyond current capabilities. It aims to enhance transit agency awareness of the range of data science methods in use by peer organizations and data analytics practices from other fields that could address public transportation challenges.

The focus of this report is especially important given the impact of the COVID-19 pandemic on public transportation. Although transit's path during and after the pandemic is in flux, technologies such as advanced sensing, new data sources, and modeling/simulation hold potential to help public transportation reinvent and transform itself after the pandemic (2).

[1] See Intelligent Transportation Systems Joint Program Office, *ITS ePrimer* (https://www.pcb.its.dot.gov/eprimer/module7.aspx) and accompanying fact sheets (https://www.pcb.its.dot.gov/factsheets/default.aspx) for overviews of AVL, APC, and AFC data.

Purpose and Scope

This report describes the state of the practice of emerging data science tools and methodologies among U.S. transit agencies. It identifies common tools and methods and common problems that advanced data science approaches can help address. The report also summarizes challenges, opportunities, issues, and potential solutions among agencies on the forefront of emerging data science. Finally, the report outlines key factors and considerations for the further adoption of emerging data science practices and tools within the U.S. domestic transit industry.

This report uses the terms "emerging" data science and "advanced" data science as broad and subjective terms with the intention of differentiating the forefront of data science approaches in transit from more conventional practices.

What is Data Science?

Data science is a field of practice involving the extraction of insights from ever-expanding volumes of data (3). Early mentions of data science as a field date to the 1960s, although observations even as recent as 2015 acknowledge that no consensus yet exists on what precisely constitutes data science (4).

Recognizing the range definitions and an array of risks that stem from this lack of agreement or consistency, Fayyad and Hamutcu (2020) proposed a working definition based on a review of academic literature and a decomposition of the most common elements—"using data to achieve specified goals by designing or applying computational methods for inference or prediction." They also further define terms embedded in this definition, as shown in Table 1-1.

Table 1-1 *Additional Terms Embedded in Data Science Definition*

Term	Definition
Achieving Specified Goals	Depending on the domain and context, can mean exploration, discovery, decision-making, prediction, optimization, or similar objectives and tasks. This is very much related to the scientific method for building knowledge from observations.
Designing or Applying	Means that activities such as designing, understanding, or examining inference methods (e.g., the study of learning from data in machine learning [ML]) or applying methods in a particular problem context (e.g., using statistical analysis or inference methods) are included as data science activities.
Computational Methods	Refers to using computers either to directly conduct a search or to aid a human in formulating or optimizing a model; for example, the search can be for a model, a structure, or an explanation and can be achieved through a variety of techniques from many fields, including analytical, statistical, or machine learning tools and techniques.
Inference or Prediction	Inference can be conducted algorithmically; this includes automated formulation of hypotheses, automated exploration of definitions of new attributes or representations, and so on. Prediction includes the cases where the goal is just to produce an optimized predictive model without necessarily gaining insights into how it works; for example, in deep learning when the goal is to achieve a certain level of performance on the predictive model output.
Data	Can be structured or unstructured. Achieving goals or getting computational inference methods to work on data often require related tasks such as data cleaning or transformation.

Source: Fayyad and Hamatcu (2020)

Although other definitions of data science exist (many of them reviewed by Fayyad and Hamutcu [2020]), this study does not aim to choose among them or propose a novel definition. Instead, the study team used this working definition as a starting point to guide its literature review and interviews with transit agencies. The team also expanded on this definition with examples of specific methods, applications, and other defining characteristics.

What are Some Conventional Uses of Transit Data?

As noted, many agencies have sought to harness their AVL, APC, and AFC data through Business Intelligence (BI) tools that convey information about the past and the present. Transit agencies have created performance scorecards and dashboards for both internal and external audiences. These data tools provide key performance indicators and summarize agency performance in areas such as safety, on-time arrivals, vehicle reliability, and other data collected by agencies. These BI tools typically communicate information using descriptive statistics such as time series plots and color-coded charts indicating whether an aspect of service is meeting an agency's performance standard.[2]

What Makes Data Science "Advanced" or "Emerging"?

The section above outlines a working definition of data science to inform the scope of this study, which also sought to identify the most progressive practices in the use of data science by U.S. transit agencies. The terms "emerging" and "advanced" are subjective, meant to characterize data science practices and methods within the transit industry, not necessarily compared to other fields. Many of the practices identified in this report may already be common in other fields, even as they are just being adopted into the transit industry today. Moreover, this report uses both terms without implying judgment as to what practices agencies *should* employ, but rather to highlight the novelty and sophistication of practices relative to one another.

This report's "Summary of Observations and Issues" summarizes transit agency input on what constitutes "emerging" or "advanced" in the context of their data science work. The project team also proposed an upfront and hypothetical set of distinguishing characteristics to guide its initial research, outlined in Table 1-2.

[2] For examples of performance scorecards, see https://www.transitchicago.com/performance/, https://www.wmata.com/about/records/scorecard/index.cfm, https://www.soundtransit.org/ride-with-us/system-performance-tracker.

Table 1-2 *Proposed Distinguishing Characteristics of "Emerging" and "Advanced" Data Science in Contrast to "Typical" or "Routine" Data Science*

"Advanced" or "Emerging"	"Typical" or "Routine"
Predictive	Descriptive
Automated processing/review/monitoring	Manual processing/review/monitoring
Learning – algorithm develops and applies rules based on data and training processes	Explicit programming – software developers build business rules based on preexisting knowledge
Higher volume and/or velocity	Lower volume and/or velocity
Unexpected purpose / exploratory	Pre-defined purpose / production
20% resources on design; 80% on data analysis	80% on design; 20% on data analysis
Majority of dollars on data analysis	Majority of dollars on hardware & software
Consider processed data as disposable	Tight data model; strict access controls
Many people with access to data	Small number of people with access to data
Dynamic models	Static models

Adapted from Guidebook for Managing Data from Emerging Technologies for Transportation *(5)*

Other Considerations

The project team outlined several additional considerations and criteria to guide its market study, including identifying examples and agencies to interview, informed by a broad review of literature on prevailing practices in data science (not confined to transit applications). Although it applied this approach loosely, the team generally considered an application or example as "of interest" or "within scope" if it aligned with two or more of the criteria, as outlined in Table 1-3.

Table 1-3 *Additional Considerations in Identifying Examples of "Emerging" Data Science*

General Consideration	Specific Questions, Criteria, Considerations
Nature of Problem Being Solved	• Is the agency trying to solve problems for which existing solutions require a lot of fine tuning or long lists of rules? • Is the agency trying to solve problems for which a traditional approach yields no good solution? • Is the agency trying to gain insights on complex problems using large amounts of data, such as sensor data or cell phone data? • Is the agency trying to streamline processes that currently require extensive manual/human effort? • Is the agency trying to identify problems/conditions in real-time?
Technology/ Methodology in Use	• Machine Learning/Deep Learning (including Supervised, Unsupervised, Reinforcement) and specific sub-methods, e.g., decision trees, random forest, support vector machines • Artificial Neural Networks/Deep Neural Networks • Data Fusion • Advanced Sensing • Digital Twins

General Consideration	Specific Questions, Criteria, Considerations
Capability Enabled	• Predicting ridership, revenue, or other performance information • Detecting fraud or anomalies in business operations • Segmenting customers to learn about current demand or future customers • Creating chatbot or personal assistant • Natural language processing or other examples of freeform text analysis • Working with speech recognition or facial recognition • Simulation and modeling • Route and Schedule Optimization
Nature and Type of Data Used	• Big/Real-Time Data (as opposed to historical/static) • Streams of data (as opposed to static repository) • Integration between multiple sources • Image/video • Unstructured text/audio • Categorical/numerical (particularly if in combination with additional characteristic above)

Limitations and Related Topics

This report focuses on opportunities and practical considerations; it does not cover more general concepts in data science and ancillary fields, including foundational advances in data collection and sharing, computing, statistics, and other technical fields; the expanding availability, affordability, and scalability of storage and cloud computing resources; and the increasing sophistication of analytical approaches that leverage them (e.g., machine learning and deep learning). The following subsection provides an overview of key concepts and terms that underly applications described later in the report as context.

Several related resources offer deeper and more detailed technical information on these topics:

- *Policy Brief: Leveraging Big Data in the Public Transportation Industry* (6)
- *Guidebook for Managing Data from Emerging Technologies for Transportation* (5)
- *Data Sharing Guidance for Public Transit Agencies – Now and in the Future* (7)
- *The Transit Analyst Toolbox: Analysis and Approaches for Reporting, Communicating, and Examining Transit Data* (8)
- *Objective-Driven Data Sharing for Transit Agencies in Mobility Partnerships* (9)
- *Big Data in Public Transportation: A Review of Sources and Methods* (10)

Summary of Key Terms and Concepts

This section summarizes key concepts and terms relevant to emerging data science in public transportation. Many of these terms do not have universally agreed-upon definitions. Although this section presents common themes and representative definitions, they may not be comprehensive. Moreover, this

section covers only terms and concepts referenced in later sections. The terms come in two broad categories:

- *Foundational Terms and Concepts* – Tools, resources, and methodologies that underly advances in data science.
- *Application Areas and Examples* – Categories of capabilities and tools enabled by many of the foundational concepts and other elements of advanced data science.

Foundational Terms and Concepts

Data

Many of the capabilities described later in this report define data broadly. Although a traditional definition of data is "characteristics or information, usually numerical, that are collected through observation," emerging data science capabilities treat data as any information represented in a numerical form, not just information conventionally *observed* in numerical form, such as survey results or experimental measurements (11). Images, both still and video, are data, as are audio records; numerical terms can describe pixel orientation/color, sound waves, and text.

Data can vary in other dimensions. Data collection can occur continuously and in real-time or in short, finite windows. They can be highly structured (e.g., measuring the on-time performance of a train) or unstructured (e.g., a collection of documents or images). They can also represent samples (e.g., a subset of passengers issued a survey) or populations (e.g., farecard timestamps from all passengers entering and exiting a subway system).

In its policy brief *Leveraging Big Data in the Public Transportation Industry*, the American Public Transportation Association (APTA) lists an array of data sources collected regularly by transit agencies, including data entered manually by agency personnel such as time tracking, absenteeism rates, and safety incidents, in addition to the APC, AVL, and AFC sources cited earlier (6).

Cloud Computing

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services). Instead of using a fixed amount of computing power or storage provided by onsite hardware (i.e., computers and servers), cloud computing enables a user to access flexible levels of computing and storage resources over a network, allowing for rapid scaling with minimal management effort or service provider interaction. These characteristics mean that users requiring significant or varying levels of computing or storage resources can purchase them through a provider instead of paying directly for the underlying assets such as hardware, software,

and networks. This is significant in the context of advanced data science methodologies, which increasingly rely on large datasets and high-performance processing. Transit agencies surveyed on their use of emerging data sources cited the importance of cloud computing resources to sharing and analyzing complex, large, and high-velocity datasets (6). That said, the cost-effectiveness of cloud computing relative to local computing and storage resources depends on the use case and agency.

Machine Learning

Computer science pioneer Arthur Samuel first defined machine learning in 1959 as "the field of study that gives computers the ability to learn without explicitly being programmed" (12). Generally considered to be a sub-domain of the field of artificial intelligence, machine learning swaps traditional computer programming—whereby software engineers write specific instructions for a computer to follow—with a process that allows the computer to program itself through experience. A computer is exposed to data and identifies patterns or makes predictions based on a model chosen and adjusted by programmers. The programmers then use a subset of the original data, set aside and not used for training, to test the accuracy of the resulting model (13).

Machine learning allows software to identify patterns in data without being programmed with explicit rules for how to detect patterns. Although the concept of machine learning has existed for decades, recent advances in data availability and the power, affordability, and scalability of computing and storage resources has accelerated its development and practical use.

Data and Multisource Information Fusion/Integration

Data and information fusion involve combining and integrating multiple sources and types of data to yield more specific or comprehensive insights (14). Fused datasets typically take on novel or new characteristics relative to the datasets used to compose them. Although data fusion originated with respect to sensor data, the concept has expanded to include a variety of data and information sources and types.

Many of the applications described later in this report integrate data from different sources. For example, a transit agency may use a combination of fare payment, vehicle location, schedule, boarding/alighting, and broader environmental (e.g., weather, traffic) data to assess demand patterns.[3] Data fusion is often a foundational step that precedes data analysis and may include aligning data along spatial and temporal dimensions, which can be difficult and time-consuming, particularly when the data come in different formats and from different sources.

[3] Although in certain fields and contexts distinctions exist between data fusion and data integration, for the purposes of this report the concepts are used interchangeably.

Data fusion can both leverage and enable machine learning. Applications with high data variety, volume, and velocity can benefit greatly from machine learning capabilities. Machine learning can also serve as a foundational tool to conduct data fusion, creating complex and extensive datasets that may challenge more conventional statistical analysis approaches (15).

Application Areas and Examples

Computer Vision

Computer vision involves extracting meaning from still or video images. Modern computer vision applications rely extensively on machine learning methods, commonly involving a process of supervised learning, whereby computer scientists and engineers train algorithms using a series of images labelled with objects or features of interest (16). Machine learning algorithms use this labelled training data to identify common features of interest, developing the ability to identify those same features in new, unlabeled images.

Computer vision has been applied in a range of fields that rely upon image-based information, including medical imaging, security, and social media. In transportation, computer vision is a central component of emerging Automated Driving Systems but also has applications in video-based infrastructure inspection, transportation operations, and security, among other areas (17).

Natural Language Processing

Natural language processing, or NLP, trains computational models to interpret, extract meaning from, and generate human language (18). Elements of NLP include identifying and determining meaningful components of words, extracting information from phrases, and retrieving or generating new information. Natural language processing underpins many common consumer-facing applications, including smart assistants, chatbots, and search engines. Professional disciplines like medicine also employ NLP, for example, to assist doctors in extracting meaningful information from vast volumes of text, such as medical records and clinical guidelines (19).

Digital Twins

The combination of advanced sensing, data fusion and integration capabilities, and analysis methodologies have enabled a new class of models called digital twins, capable of producing a detailed, real-time picture of an entire system's performance and conduction. Digital twins combine real-time and historic data from available sensors, transducers, communication systems, processors, and remote terminals that are increasingly built into transit assets. Technology providers suggest that, by allowing for real-time modeling of asset condition and performance, digital twins may enable more effective and efficient asset maintenance and repairs while minimizing equipment or route downtime.

Vendors have begun to offer transit-focused digital twin systems; however, the use of digital twins in transit remains in an emerging stage. In-use examples in the literature were largely limited to transportation planning applications, although numerous sources also identified the potential for digital twins to assist in operating and maintaining rail networks (20, 21, 22, 23).

Section 2 Methodology

Literature Review

The project team reviewed the literature to identify emerging and advanced data science methods and practices that could address the needs of transit agencies. The review included technologies and methods both within and outside the transit industry. It emphasized applied techniques and included examples both developed in-house by transit agencies and developed by contractors, vendors, and other private partners and academia.

The review began with a survey of the Transportation Research Board (TRB) Transportation International Documentation (TRID) database, using an initial set of key search terms and select committees of interest (Table 2-1 and Table 2-2). The project team then expanded its list of key search terms from the initial set. The team also reviewed materials from recent conferences and workshops, as well as project websites of recent recipients of innovation-oriented FTA grants (e.g., Mobility on Demand [MOD], Integrated Mobility Integration [IMI], and Accelerating Innovative Mobility [AIM]).

The project team performed a general internet search for examples of advanced data science in transit using the previously identified search terms. Sources included news articles, press releases, presentations, transit research groups, and industry reports. The team conducted an additional internet search, specifically targeting examples of advanced data science outside of transit that could potentially benefit transit agencies. This search included examples from road, rail, aviation, freight, and private industry.

Table 2-1 *TRB Committee Materials Reviewed*

Transit Research Analysis Committee
Artificial Intelligence (AI) and Advanced Computing
Emerging and Innovative Public Transport and Technologies
Intelligent Transportation Systems
Information Systems and Technology
Standing Committee on Transit Data
Standing Committee on Bus Transit Systems
Standing Committee on Transit Management and Performance
Standing Committee on Rural, Intercity Bus, and Specialized Transportation

Table 2-2 *Key Search Terms*

Key Search Terms	
• Artificial Intelligence • Machine Learning • Decision Trees • Decision Making • Mathematical Models • Random Forest • Decision Support Systems • Case Studies • Support Vector Machines • Data Analysis • Payment • Linear/Logistic Regression • Text/Textual Analysis • Detecting Fraud/Anomalies	• Clustering Algorithms • Natural Language Processing • Segmenting (Customers/Riders/Users) • K-Means • Flagging Content • Demand • Neural Networks • Predictive Analytics • Chatbot/Personal Assistant • Digital Twins • Predicting (Ridership/Revenue Performance) • Classification Algorithms

Interviews

The team interviewed two separate groups of transit agencies using separate questions. The first group comprised seven transit agencies and one partnering transit lab and focused on agencies conducting innovative data science work (see Table 2-3 for overview of agency characteristics). Through these interviews, the team learned about these agencies' current and planned activities and observed factors affecting their ability to invest in and adopt emerging data science capabilities. Agencies in this group generally covered medium to large metropolitan areas in different regions across the United States. The team identified these agencies through a combination of existing FTA grant awards, participation in relevant TRB committees and APTA projects, examples identified from the literature, and recommendations from prior interviews.

Table 2-3 *Range of Agency and Service Characteristics Covered Through Round 1 Interviews*

Characteristic	Range Covered Through Interviews
Service Types	Heavy Rail, Light Rail, Bus, Demand Response, Bus Rapid Transit, Ferry, Commuter Rail, Hybrid Rail
Annual Passenger Miles of Travel (2019)	Over 300 million to multiple billions
Annual Unlinked Passenger Trips (2019)	Over 70 million to multiple billions
Service Area Population (2019)	Between 1 and 10 million
FTA Regions	6 FTA regions represented

Characteristic	Range Covered Through Interviews
Topics Covered	• Interviewee definition of emerging data science • Relevant project/application examples, including: - Data used - Tools/techniques leveraged - Project origins, motivations - Goals/expected outcomes of project/application • Resources, partners, and expertise leveraged • Interactions with other agencies (to learn and/or share) • Future work in emerging data science • Recommendations for further research/interviews

The team also conducted discussions with four agencies that expressed an interest in emerging data science practices through prior innovative work, using a separate set of a discussion questions (see Table 2-4).

Table 2-4 *Range of Agency and Service Characteristics Covered Through Round 2 Interviews*

Characteristic	Range Covered Through Interviews
Service Types	Demand Response, Bus, Bus Rapid Transit, Vanpool, Ferry, Streetcar, Trolleybus
Annual Passenger Miles of Travel (2019)	Between 10 and 800 million
Annual Unlinked Passenger Trips (2019)	Between 3 and 200 million
Service Area Population (2019)	Between 200,000 and 2.5 million
FTA Regions	4 FTA regions represented
Topics Covered	• Interviewee definition of emerging data science • Potential areas for emerging data science to benefit agency/areas of interest • Sources of information, guidance, inspiration, new ideas for investing in data science capabilities • Overview of in-house data capabilities/expertise • Experience working with data science-focused vendors • Most significant factor influencing agency investments in data science • Information or resources that would be helpful

The project team received guidance from FTA staff in selecting the agencies interviewed in both groups. The project team agreed not to identify the agencies interviewed or associate them with their observations to encourage candid dialogue. **Any references to specific agencies, examples, or systems in this report are based on literature review findings and not the interviews.**

Section 3 Key Findings

Transit Applications in Use Today

Through its interviews and literature review, the project team identified a range of emerging data science practices and applications that transit agencies are using; this section summarizes several of the more prevalent examples. The team found a wide variety of emerging data science applications using AVL, APC, AFC, vehicle diagnostics, Location-Based Services (LBS) data, and dispatch and operations data. However, it also identified a gap between an extensive body of hypothetical examples presented in the academic literature, particularly examples of machine learning and artificial intelligence applied to the transit domain and more limited real-world applications. This gap suggests that academia remains a robust source of new ideas and methodologies. However, academic proposals and partnerships geared towards the practical realities of U.S. transit agencies could improve technology transfer.

Asset Health Monitoring and Predictive Maintenance

Several agencies interviewed are either using or planning to use emerging technology and machine learning algorithms to monitor the health of their assets. At the time this research was conducted (2020–2021), the technology appears to be limited to vehicle maintenance; however, it is plausible that this scope will expand to include other transit assets (24). Several domestic transit agencies mentioned their adoption of software that uses existing telematics (an interdisciplinary field that encompasses telecommunications, vehicular technologies, electrical engineering, and computer science) along with machine learning to identify when a vehicle is likely to need maintenance. The tool monitors emissions, accelerometer and location data, and previous repair histories to estimate remaining time before a repair is needed. A dashboard provides advance warnings, aiding agencies in scheduling maintenance, improving efficiency, and helping to prevent vehicle breakdowns (25, 26). APTA cites similar examples of transit agencies leveraging predictive modeling to anticipate bus breakdowns (6). The Utah Transit Authority also reports using radio-frequency identification (RFID) tags to monitor and predict maintenance needs on major components on its railcars. These tags enable the agency to gather data on how long and how frequently these vehicles and components are in operation (27).

Several additional examples of emerging asset health monitoring techniques exist outside of transit but are potentially transferrable. Many of these examples come from heavy rail, which shares characteristics with transit and could serve as a testing ground for new and emerging technologies. For example, the Federal Railroad Administration (FRA) has sponsored research into using computer vision-based approaches to automate rail inspections (28).

Additionally, private sector companies have for some time leveraged advanced sensing and machine learning to predict vehicle maintenance needs, including in delivery vehicles and airplanes (29, 30, 31, 32). The concept of digital twins, discussed earlier, has been particularly emergent with respect to vehicle and transportation system maintenance, including as applied to commercial fleet vehicles and aircraft.

Occupant Counting and Monitoring

Transit agencies are using various methods—some emerging or advanced—to monitor ridership and rider behaviors. Transit agencies have traditionally used manual methods for measuring ridership to conduct route planning, reporting, and performance measurement. Now, automated and real-time sources offer new opportunities, including tracking ridership changes, predicting demand, and ensuring passenger safety. Real-time onboard crowding information has become particularly useful during the COVID-19 pandemic. For example, agencies have adapted onboard camera systems to automatically track ridership levels and determine whether riders are social distancing during the COVID-19 pandemic (33).

Several of the interviewed agencies mentioned obtaining accurate passenger counts as a challenge, both for performance measurement and for communicating crowding conditions to passengers; agencies cited a range of factors, including accuracy and reliability of APCs. One of the agencies interviewed has applied a machine learning algorithm to onboard video footage to verify the accuracy of their APC systems. The agency uses these data to produce regular ridership updates. Other similar examples are available in the literature, including Miami-Dade County Transit's use of a computer vision-based system to monitor for passenger crowding and the Metropolitan Transportation Authority Long Island Rail Road's use of APCs, train car weight sensors, and computer vision-enabled cameras (34, 35).

Emerging data science approaches have also enabled agencies to better understand passenger movements and crowding in subway stations. One example in the literature describes integrating rail network and station circulation simulation models with vehicle location, passenger boarding, and fare payment data to regularly monitor station crowding conditions (36). One agency interviewed partnered with a private vendor to deploy sensors in two of its stations to augment and validate existing measurements of real-time crowding and origin-destination flows within stations. However, the agency noted challenges in using the resulting data for certain purposes. Although the system was helpful for understanding passenger movements within stations, it was less effective in measuring activities such as fare evasion. The agency noted that the system would likely be cost-prohibitive to deploy fully across

other stations, but it identified more targeted uses such as assisting in station redesigns.

Operational Tools

Several transit agencies are using emerging data science methods and tools to improve their day-to-day operations, particularly to improve service reliability and transparency to both customers and personnel. Whereas some of these examples involve the use of advanced data science methods (e.g., machine learning), others exhibit more fundamental approaches (e.g., data integration and sharing) that are noteworthy for the capabilities they enable.

Bus Arrival Prediction – One transit agency described its efforts to better predict bus arrival times with machine learning, leveraging the availability of General Transit Feed Specification (GTFS) real-time data and other sources. Its modelling effort considers historic data on travel times between stops and data on environmental conditions such as traffic and weather. The project is being developed in-house, based on the work of an international peer agency (37).

Operations and Dispatch Tools – Several agencies mentioned examples of operational tools that integrate data from multiple sources into a combined platform, facilitating easier and faster access to and sharing of information and, in some cases, novel optimization tools. Agencies described these examples as "emerging" because of the non-trivial nature of integrating multiple sources of historic and real-time data and presenting those data to users in an intuitive manner. For example, one agency described its mobile dispatch application that integrated and presented data to bus dispatchers in a user-friendly format, allowing it to replace paper-based processes. In discussing this application, one agency representative commented that "What is advanced about [our agency's] use of data science is not the technical methods used, but what occurs around the data science," and "…advanced prediction algorithms are not valuable if they present information in ways that are not useful to the users."

Many private dashboard tools also advertise fleet management capabilities, including automated routing, demand prediction, and real-time information sharing. For example, Oahu's TheBus uses a private real-time vehicle monitoring platform that enables the agency to manage headway between buses. The system uses real-time traffic and vehicle location data to determine whether an operator is too close to the vehicle ahead or if it is moving too slowly, adjusting the operator's layover countdown timer accordingly (38).

Operator/Assignment Ratio Optimization – Transit agencies must schedule "extraboard" operators who can fill in for their absent colleagues to maintain service levels when operators experience unexpected absences (e.g., for sick leave). Managing extraboard staffing entails a balance between covering scheduled service and minimizing cost (i.e., avoiding overstaffing), all within

prevailing legal and contractual constraints, including collective bargaining arrangements (39). One agency indicated that it is working with a vendor to explore the use of machine learning to predict absence rates, using historical data to optimize its operator/assignment ratios and extraboard staffing; this example mirrors others cited in available literature (6, 40).

Customer Service/Interaction – Several agencies provided examples of how they are using emerging data science tools to communicate with customers. One agency indicated that it is beginning to use natural language processing techniques to better understand customer sentiment in freeform text from surveys, customer service messages, and social media. Another mentioned partnering with a technology vendor to develop a real-time communications platform that not only allows for interactions with customers but also data collection that should help reveal patterns in customer inquiries and complaints (e.g., routes and locations on routes that commonly correspond to complaints). Although in its early stages, Amtrak has applied NLP methods to help customers book travel through the phone (41).

Safety and Security – The literature contains several examples of emerging security technologies applied in airports that make use of sensors similar to those used by some transit agencies and may be transferrable (42). Machine learning and computer vision technologies are also being applied in trucking and heavy rail operations to help monitor for driver fatigue and distraction (43). Although the emerging technologies being applied today focus on ensuring passenger and personnel safety, it is important to note the privacy implications of this technology.

Planning, Scheduling, and Performance Management Tools

Service planning and scheduling are inherently data-driven processes, and many examples of emerging data science practices focused on this area. Several agencies mentioned the use of LBS data to augment traditional data sources and reveal nuances in travel demand patterns. Agencies suggested that these data resources and accompanying tools that some vendors (e.g., Streetlight) provide have proven particularly useful during the COVID-19 pandemic. For example, one agency indicated that this new data source helped to reveal travel demand outside of typical commuting peaks, which became more prominent and critical as many office workers shifted to remote work. Other agencies commented on their use of LBS data to identify mode-specific travel volumes along certain corridors, pointing to the use of machine learning by the vendor to predict the travel mode of each data source based on trip characteristics (e.g., travel speed and speed profile, origin and destination) (44). LBS data has the potential to give transit agencies insight into travel demand in areas not currently served by the agency's fixed-route service.

In addition to using planning and scheduling data and tools, one agency integrated data sources to determine on-time performance. This agency commented that, in many respects, their most "advanced" work in data was not to develop and deploy novel and emerging analysis methodologies (for which they had ample in-house staff and expertise), but to integrate and combine datasets that were developed at different times, with different vendors, and for different purposes. For example, staff described the extensive integration work required to accurately measure on-time performance and the impact of dropped trips, citing challenges rooted in hardware reliability and compatibility, software flexibility, and data processing. The agency ultimately developed a solution that combined new hardware, advanced data processing, and an option for manual review to improve the completeness and accuracy of its on-time performance data.

Summary of Observations and Issues

Transit agency representatives interviewed offered observations on a range of topics, including the types of applications and other data science work emerging in their agency, the motivations behind investments in these new capabilities, the factors that have influenced success, and challenges in developing and deploying new data resources. This section summarizes common themes.

Overall

The transit sector appears to be transitioning from a push to collect high-volume and high-velocity data to managing, integrating, and ultimately conducting advanced analyses on it. Many agencies described their work in recent years to collect vehicle location, passenger count, and fare collection data as a precursor to advanced data science work. Several discussed how these data sources were originally established to support reporting and were separate from one another and other data resources. Some agencies described their efforts to integrate disparate data into more comprehensive and complex datasets. Several agency representatives commented that, due to a variety of factors—including patchworks of vendor software and hardware, often implemented during different time periods without consideration for how they might later be integrated with one another—data integration was a far more complex task than leveraging any new or emerging analysis approaches (e.g., machine learning or deep learning). Many also cited the importance of data hygiene (i.e., the process of ensuring that data is free from errors and missing information) in enabling data integration and more sophisticated analysis approaches.

Definitions of what constitutes advanced or emerging data science vary, but with common themes. As noted, emerging data science is not necessarily a formally defined term or concept. Each interview began with a discussion on

how agency representatives characterized emerging data science and how they distinguished between "emerging" data science and data science methods and applications that are more routine or commonplace. Perspectives varied widely but centered around several common themes, outlined in Table 3-1.

Table 3-1 *Observations from Interviewees on Defining Advanced and Emerging Data Science in Transit*

Theme	Relevant Observations and Definitions from Interviews
Data Integration	• Combining disparate data sources to yield novel insights. • Business tools that integrate data and information. • Data science is advanced when data are combined across a large number of enterprise systems or when data cover longer timeframes. • Collecting and organizing data in a way that may not have been possible due to prior storage or database technology constraints. Using data to generate better statistics for evaluation. Gathering different data streams is possible using machine learning and can allow transit agencies to make improvements to services and operations. • Good data hygiene is an absolute prerequisite of advanced data science and analytics.
Focus on Usability	• Turning data into information that is usable to an end consumer. • Advanced data science is when people go through the effort of making data usable in simpler ways. • What can make data science advanced is not the technical methods used but what occurs around the data science. For example, advanced prediction algorithms are not valuable if they present information in ways that are not useful to the users. Usability and presenting data in effective and deliberate ways is critical for advanced data science.
Problem-Driven, In-Depth Analysis	• Being confronted with a data source (a database or stream of information) and being able to look at the data in its rawest form to manage, combine, reshape, clean, and visualize this data before analyzing it. • Applying the scientific method to data to solve a problem. Looking at the data without a specific problem to solve is just data exploration. • Data science ranges from statistical analysis to machine learning and provides actionable information to reach agency goals. Simple summary statistics can provide a lot of benefits, but once you get beyond this and look deeper into the data then this what we would consider advanced.

Academic examples of how machine learning may benefit transit planning and operations are plentiful, but instances of in-use applications are more limited. The academic literature includes many theoretical examples of how data available from transit agencies could support increasingly sophisticated analysis. Example application areas include (but are not limited to) the following:

- *Optimizing vehicle and operator assignments* – For example, using reinforcement learning approaches to optimize the assignment of vehicles to routes around performance, service quality, and cost (45).

- *Predicting passenger flows and behaviors* – For example, using deep learning and convolutional neural networks to improve origin-destination and boarding-alighting predictions for both bus and train lines (46, 47, 48).
- *Predicting and mitigating the impacts from unplanned service disruptions* – For example, demonstrating how natural language processing techniques can be used to automatically extract information on delays, impacts, mitigation strategies, underlying incident causes, and insights related to potential actions and causes, all from unstructured text in written reports (49).
- *Predicting travel times* – For example, using AVL and APC data to train artificial neural networks to predict bus travel times in Washington, DC (50).

Although many of these academic examples may influence and inform what is adopted, examples of practical, on-the-ground applications are more limited. Some universities, however, operate transit-focused labs that partner with transit agencies to develop and test novel analytical approaches. Examples cited by interviewees include labs at MIT, The Ohio State University, the University of Minnesota, and the University of Washington. In addition to their value in conducting research, interviewees from several agencies indicated that university research partnerships were valuable in helping to develop a hiring pipeline.

Data Sourcing, Collection, and Management

Agencies described common sources of data informing their emerging data science work, including traditional sources such as surveys, newer in-house sources such as APCs, and novel third-party sources such as location-based services. Many agencies mentioned that new data sources are key to developing advanced insights into rider needs and behaviors. As noted above, interviewees commonly cited APC, AVL, and AFC data as critical in current advanced analysis work or expected to play a more critical role once fully operational and/or integrated with other datasets. Agencies also discussed new third-party and vendor data sources, including LBS data (i.e., data collected through smartphones and other devices equipped with cellular connectivity and global positioning system [GPS] receivers), advanced sensors, video data, and onboard vehicle diagnostic data. Several agencies emphasized that novel data sources are not inherently useful without being aligned with an agency's priorities and resources.

Agencies described three general categories of data informing their emerging data science work:

- Legacy/Traditional Transit Data Sources – traditional passenger counts, ridership surveys

- Emerging/Recently Emerged Internal Data Sources – AVL, APC, AFC
- Novel Third-Party Data Sources – LBS, sensor/camera data, private sector mobility data (e.g., HERE, INRIX, Uber Mobility, Google)

Data hygiene is a precursor to exploring more advanced analysis capabilities. Interviewees stressed the importance of quality data and data cleaning, citing the amount of time it can take to clean the data as a major impediment but also a critical prerequisite to using emerging data science. One agency representative commented that "in all cases, most of the work is on the data preparation stage; this isn't to downplay new technologies or methods, but we do spend most of our time at the beginning."

Interviews reinforced the importance of data quality and organization in an agency's readiness to apply newer analysis approaches. Other explorations of data practices among transit agencies have highlighted similar issues, suggesting that although agencies are fairly consistent in collecting APC, AVL, and AFC data, data quality lags due to a "broad spectrum of hardware, software, and internal business practices" (51). Another study observed that transit agencies have not adopted data governance approaches very widely, and among the few agencies that have, all began quite recently and for the same reasons: to improve data quality and ensure consistency of data and analysis results (8).

Several agencies stressed the importance of keeping data open source to encourage information exchange and develop emerging practices. Publicly-available data also make it easier for agencies to partner with research institutions and transit labs. Some agencies indicated that academic data sharing arrangements yield useful research results and help to develop a more robust pipeline of potential staff hires with both data and transit experience.

A recent report published by the National Academies of Sciences, Engineering, and Medicine through the Transit Cooperative Research Program (TCRP) provides more detailed information and guidance on data sharing for public transit agencies (7).

Agencies described two general models for approaching new "advanced" analysis opportunities. Agencies developed new data science applications and resources in one of two ways:

- *Exploratory studies of existing data sources* – Many agencies fused and analyzed AVL, APC, and AFC data to pursue new insights. Interviewees described the "advanced/emerging" aspect of this work as preparation and integration required to conduct analysis across multiple datasets. These examples seemed to be more opportunity-driven—i.e., agencies seeking to extract latent value from datasets already available to them.

- *Problem-driven, standalone data collection and analysis efforts* – Several agencies described specific projects involving dedicated/focused data collection combined with advanced analysis. Many of the explicit examples of the application of AI/ML techniques fell into this category, including involving computer vision and predictive analysis for bus maintenance. These examples seemed more problem-driven—i.e., agencies defined a problem and identified an advanced analysis method (sometimes combined with the collection of new data) as a potential solution.

Both models are valuable, and some interviewees described using both approaches. Exploratory studies tended to involve (or even require) in-house staff with relevant expertise in both data and transit; these staff need to understand the details of the data available and identify opportunities for available data to address an organizational or operational need. Problem-driven studies tended to use vendors or consultants who deployed targeted capabilities to address a priority need, although staff expertise was no less important in these instances (see "Feasibility Analysis" for further discussion on the importance of staff capacity).

Integration of existing systems can present challenges and limitations. Several agencies described challenges of integrating existing datasets and/or leveraging new analysis methodologies (e.g., machine learning) on existing data. In some cases, agencies described experiences with vendors or consultants who approached them enthusiastically about new analysis, only to be disappointed once the agencies saw the results produced on the actual data. Agencies generally cited the state of their existing data (and not vendor or consultant capabilities) as the impediment. They also emphasized the importance of vendors and consultants being familiar with the state of their data and the transit domain. Vendors' lack of expertise in the transit domain was sometimes mentioned as an impediment.

Analysis Tools, Methodologies, and Use Cases

Agencies can use machine learning to address targeted needs, but advanced data science methodologies are not a panacea. Interviews and the literature review revealed a limited number of deployed examples of advanced analysis tools rooted in machine learning. The examples identified reinforced the notion that the current state of practical machine learning applications remains limited to specific, targeted, and narrowly defined problems. Examples of machine learning being used by transit agencies appeared to follow two general patterns:

- *Limited in-house use of machine learning for application development and experimental/exploratory analysis* – This included the use of machine learning to predict bus arrival times, perform customer segmentation and

ridership predictions, and conduct sentiment analysis on free response survey data.

- *AI and machine learning embedded in vendor systems* – Several transit agencies described their use of vendor systems that leveraged machine learning as part of pilots and, in some cases, fully operational applications, including predictive maintenance applications, computer vision-based passenger counting systems, mobility data platforms, and business intelligence platforms.

Approaches that are "advanced" due to the nature of data being used are valuable. When asked to describe what constituted "advanced" data science to them and to offer project or application examples, several agencies focused on new ways of integrating data, combined with more common analysis methods, as the advanced aspect of their work. Although emerging and novel analysis methods such as machine learning and deep learning may be percolating in other sectors and in the transit-focused academic literature, agency representatives approached the terms "emerging" and "advanced" as relative and industry-specific.

Institutional Issues and Factors

Transit-specific domain knowledge and technical expertise in data are important. Many of the agencies interviewed emphasized the importance of staff with specific expertise and experience. Agencies regularly tied success to individuals or teams with knowledge of both data/analysis tools and the agency/transit sector. Agencies also seemed to weight technical and domain knowledge equally. It is critical for the data scientist to understand transit data, which can be complex and non-intuitive to new users. Agency representatives suggested that the most successful projects heavily engaged transit staff with knowledge of the transit data. Agencies also suggested that familiarity with the transit sector seemed as important as experience and expertise in data science, an observation that applied to in-house staff as well as vendors and consultants.

Agencies also cited challenges in recruiting and retaining staff with data expertise and skillsets that are valued across a range of industries. Previous studies also cited resource and institutional constraints, suggesting that frustrating conditions for accessing and sharing data across organizational units can impact retention (51).

Interviewees highlighted several successful approaches for recruiting and retaining skilled data scientists Several indicated that university partnerships offered a pipeline for hiring early career staff; when these partnerships include hands-on research projects, recent graduates bring existing agency knowledge in addition to data expertise. One agency also described success in recruiting data scientists from scientific fields (e.g., biology, physics, and ecology) in which data science is a tool but not the core focus. Agency representatives also

pointed to the nature of the work offered as a factor in retention, suggesting that offering the ability to conduct self-directed, exploratory data work – in addition to need-driven projects – was appealing.

Locating data science work within the organizational structure is important. Although the teams interviewed resided in different areas of their respective organizations, they shared generally consistent insights into organizational factors behind their success and/or challenges:

- *Various organizational locations* – The larger data-oriented teams interviewed resided in different locations, including in information technology (IT), planning, and technology/innovation departments. Within several of the agencies interviewed, data-focused staff and teams reported to agency leadership, for example its chief of staff, alongside teams focused on other cross-cutting priority functions (e.g., equity and civil rights). Agency representatives suggested that this prominence was particularly important for small teams (or even individuals) tasked with data innovation within smaller agencies.
- *Dedicated responsibility for data* – Agencies indicated that it was important that responsibility for data not fall solely to an operational unit primarily charged with service delivery. Although the agency representatives interviewed emphasized the importance of operational teams as partners in data collection, interpretation, analysis, and end use, they also stressed that the service delivery would (necessarily) always take priority over data work if housed in the same organizational unit.
- *Agency-wide reach and value* – Agency representatives interviewed emphasized the importance of data as a cross-departmental resource and suggested that the organizational location of these teams should reflect that characteristic, regardless of the precise positioning within an agency. Moreover, some individuals remarked that titles and reporting structure could help to emphasize the importance of data, both internally and externally.

Different models should be used for accessing advanced capabilities. Agency representatives described their experience in training, hiring, and retaining in-house staff to conduct advanced data work and in obtaining services and expertise from vendors and consultants. Both models present potential challenges and risks.

Agencies that leverage in-house staff appreciated the control they had over data, projects, outputs, and outcomes. Some agency representatives suggested that it was difficult to imagine contracting out work that required deep understanding of their agency, including the state of its data and data systems, its organizational culture, and its needs. One agency representative stressed the value that even a single data scientist with a broad skill base (ranging from data

collection and management to visualization and user-interface design) could offer.

Hiring dedicated in-house staff, however, may not be an option for all agencies and, even for those that can, hiring and retaining employees with highly valued skillsets is challenging. Procuring services or products from consultants and vendors offers agencies an alternative that avoids some of the drawbacks associated with hiring staff in-house. However, agency staff also described some risks of relying on vendors and consultants, including vendor-lock in and more limited control over project outcomes or data formats. Individuals interviewed emphasized that vetting of vendors and vendor capabilities and experiences with transit data and experience with applications related to transit is critical. Agencies also suggested that "data readiness" was important for successful vendor-led projects, or at least for the vendor to have an awareness of the relevant data and its condition. Some agencies described experiences in which a vendor had difficulty delivering on promised results due to the state of the agency's data.

Several agencies also described accessing expertise and new capabilities through university research partnerships. These partnerships not only offered transit agencies access to academic researchers, but also helped to educate students on the practical realities of transit and establish a pipeline for recruiting students who had developed both transit and data expertise.

Use varying mechanisms for learning about new capabilities. Interviewees described several avenues for learning about new and emerging data science tools and capabilities:

- *FTA/TCRP research* – Many agencies look to FTA and TCRP for guidance and information on common and emerging data practices.
- *TRB and APTA publications, and participation in meetings, committees* – Many interviewees recommended participation in the TRB Standing Committee on Transit Data (AP090) and found its activities valuable for both learning about new capabilities and connecting with peers in other agencies (52).
- *Hiring of interns* – Several agency representatives traced new and emerging data capabilities to interns hired from relevant transit- and/or data-focused programs. In some cases, interns developed proof-of-concept tools that were later developed into production systems. In others, agencies hired interns as full-time staff upon the completion of their degree; with a unique combination of academic-based data expertise and transit agency experience, former interns represented compelling job candidates.
- *Informal peer groups* – Several agency representatives indicated that they maintain informal networks of their peers in similar agencies, using these groups to not only learn about new capabilities, but also troubleshoot

issues with existing systems or tools under development. Some individuals even described conducting informal residencies or exchanges with peer agencies to learn more about how they configure and deploy new tools.

- *Vendor-initiated engagement* – Smaller agencies, in particular, noted that they learn about new capabilities through vendor presentations. Some participated as beta testers for startups or established companies developing new tools, gaining access to the company's tools and capabilities for free in exchange for feedback. While these agencies did not continue as paid users of all services, tools, or vendors, they indicated that it offered a useful trial period and they ultimately decided to adopt some tools trialed in this manner.
- *Self-directed learning* – Some agency staff acknowledged the value of both curiosity and extensive online resources in key areas. While not an avenue that addresses all needs, particularly related to specialized transit-specific tools, extensive online user communities exist for common tools like R and SQL.

Section 4

Feasibility Analysis

This section provides recommendations and other information for transit agencies to consider on their journey towards implementing advanced data science methods and tools.

Opportunities

Explore opportunities to unlock latent value in datasets that are already available. Transit systems and agencies produce extensive data, even in the absence of some of the newer systems mentioned earlier in this report (e.g., APCs and AVL). Integrating and exploring available datasets may offer as much as or more value than new data collection efforts or novel analysis approaches. Several agency staff indicated that working across their organization to identify opportunities for existing datasets to address needs was a significant part—and, in some cases, a majority—of their role. Moreover, one agency representative noted that this self-directed, opportunistic, and exploratory work was a significant factor in retaining staff with data science expertise.

Consider using new analysis methods to address narrowly defined problems. Emerging data sources and analysis methods offer potential benefits and can complement or augment existing data. However, it is important to explore new capabilities and applications in a problem-driven or use-case-driven manner rather than a solution-driven approach. For example, many of the use cases described above involving the use of machine learning addressed a specific problem or need—improving the predictability of bus arrivals, reducing the schedule impacts and costs associated with unexpected bus breakdowns, and measuring crowding on buses and trains or in stations.

Increasing commonality of APC, AFC, and AVL data pose an opportunity; standards can help advance analysis opportunities. The increasing prevalence of data from APC, AFC, and AVL systems present an opportunity to measure, analyze, and understand transit usage patterns more comprehensively and precisely, particularly when integrated with one another and other data sources. A current TCRP project will identify opportunities to make these data easier for transit agencies to access and manage, building on the success of the GTFS for standardized schedule data (53). This work, combined with the increasing availability of these data sources, should create an environment conducive to the development and availability of open source and adaptable analysis tools. Transit agencies should consider looking for and using common data formats including formal and informal standards. Common data formats and standards can improve the re-usability of data analysis tools and programs and improve knowledge transfer between agencies.

Risks

Data integration and unused datasets. Data should be collected with a strategic view toward use and integration. Collecting data opportunistically may result in disparate data sources that are challenging to integrate later. Several agencies described past experiences in deploying data collection systems/hardware at different points in time and the ensuing challenges of trying to integrate them subsequently. In one case, an agency described installing redundant hardware to compensate for data integration challenges.

In approaching data pilots and use cases, agencies should consider focusing on applications that address a clear and evident business need, address leadership or organization pain points, or help the agency meet its goals (5).

Vendor lock-in. Vendors can serve as a critical resource, particularly for smaller agencies, but transparency into analysis approaches, ownership of data and results, and misalignment between expectations and actual capabilities can present risks. Interoperability requirements can help to mitigate this risk and can simultaneously avoid risks associated with large systems being out of date upon deployment (51). As discussed below, targeted and strategic use of vendors accompanied by realistic expectations can further mitigate risk.

Unanticipated or malicious uses. Agencies should consider risks associated with intentional malicious use and unintentional misuse of new data resources and tools. Potential issues include (but are not limited to) the following:

- *Bias and fairness* – The expanding use of data presents opportunities for biases to be reinforced if not recognized. Novel data sources and collection methods could over- or under-represent certain groups, while new analysis approaches could make bias more difficult to observe. For example, because machine learning typically relies on historic data, in some cases annotated by humans, and involves probabilistic approaches that are difficult to audit, its application raises potential bias and fairness concerns. That is, if an algorithm is trained on a dataset that contains biases or represents biased conditions, then the resulting algorithm and its use to inform decisions may perpetuate or amplify these biases.
- *Privacy* – Emerging data science tools, practices, and resources present potential privacy concerns, particularly if they involve the collection or use of sensitive data such as images or locations. Even datasets that have been anonymized due to the sensitive information they contain can be deanonymized and, while data providers may institute consent and use agreements to help protect privacy, datasets may pass through multiple users or custodians with or without these protections intact (54).
- *Unanticipated uses* – Transit agencies should consider how tools used for prediction could be repurposed in unexpected and unplanned ways

leading to unintended consequences from the data science work. For example, data used to predict employee performance metrics may help an organization to optimize a team but could raise significant concerns if used to influence hiring and human resources decisions.

Challenges and Limitations

Availability, quality, and interoperability of existing data sources. As noted elsewhere in this report, data quality and interoperability present more significant challenges than data availability or the sophistication of analysis approaches. Even more basic, some newer data collection systems pose reliability challenges and may not fully or seamlessly integrate with existing systems. Agencies should consider how new data sources might integrate with existing datasets, which can unlock value, even in the absence of emerging analysis approaches (e.g., those involving machine learning).

Available in-house expertise. Nearly all the agencies interviewed for this project cited the importance of in-house expertise for their success in developing and applying emerging data science approaches, methods, and tools. Some of these agencies had sufficient budget and leadership support to build internal data science teams, but smaller agencies that relied more upon vendors and consultants highlighted the importance of having one or two individuals on staff with data knowledge and experience to guide and oversee projects.

End-user adoption. Agencies should consider from the start the needs and potential concerns of end users, particularly if they fall outside of core data teams or staff. While not all the projects and data work described by agency representatives had end users beyond the teams responsible for data, many were intended for operational business units and frontline personnel who expressed skepticism or reluctance to adopt new tools into their work. Several agencies indicated the importance of working upfront with end users to understand their needs and concerns. They also described how showcasing initial capabilities has helped to convince skeptical users of the value of new tools.

Resource Needs and Adoption Considerations

Consider data hygiene as a precursor to exploring more advanced analysis capabilities and approach new data collections intentionally and strategically. Data quality and organization are critical to success, even in developing and deploying narrowly focused applications. Improving the quality and interoperability of existing datasets is an example of advanced data science in transit that can represent a valuable initial step toward pursuing advanced capabilities. In addition, considering how new datasets will be managed, maintained, and integrated with other data sources is critical as agencies look

to expand their data science footprint. Simply collecting and storing data is not enough to make data useful.

Guidebook for Managing Data from Emerging Technologies for Transportation provides more detailed guidance for agencies looking to expand their use of emerging data sources (5).

Domain and technical expertise are both critical. Not all transit agencies will have the resources to hire dedicated teams of data scientists, and even for those that do, recruitment and retention can be a challenge. However, a single staff person with relevant experience and domain expertise can be critical to successfully guiding and overseeing projects. Several agencies suggested that a single person who can oversee the full data lifecycle—and ideally establish a plan or architecture around the entire data pipeline—can be sufficient. They emphasized that, particularly for agencies with limited capacity, a generalist who has an understanding of the full data pipeline—from collection and ingestion, preparation, warehousing, computation, and presentation/visualization—can be hugely valuable.

Moreover, among more expansive data teams, expertise in both data and transit as an application domain are, at least in the aggregate, equally important. Agency representatives interviewed for this project seemed split on which to focus on as a starting point in hiring. Some believed that individuals with deep and broad data expertise could offer the most value and would ultimately learn how to apply that knowledge in the context of a transit agency. Others believed that although some baseline data knowledge was important, transit agencies can present a unique environment in which to apply data science and saw limited potential to translate experience from other fields or industries. Instead, they observed great value in individuals who already had familiarity with data in the context of transit and/or the dynamics of a specific agency, even if they had to rely more on consultants, peers in other agencies, or on-the-job learning to develop more general data science knowledge and expertise.

Use vendors strategically, with realistic expectations. Vendors and consultants can offer great value, particularly in applying emerging practices and for agencies with fewer in-house resources. For transit agencies and vendors/consultants alike, approaching new data-focused projects with realistic expectations is critical for success. Agencies described disappointing results when approached by vendors that could offer sophisticated analysis capabilities but had little familiarity with the agency's data resources and the quality of those data resources. Agencies described more positive experiences working with vendors on targeted, narrowly-focused, and strategically-deployed capabilities. This could include systems where vendors collect necessary data directly and/or tap into standardized data interfaces rather than proprietary agency data repositories (e.g., bus predictive maintenance tools that leverage

telematics data available through a standard Controller Area Network (CAN) bus connection) or deployments of an emerging data science tool to validate, rather than replace, legacy data collection methods (e.g., using computer vision tools to validate passenger counts).

Leverage informal peer networks. Interviews reinforced the uniqueness of specific transit agencies but also commonalities in their collection, management, use, and sharing of data, the challenges they face in these areas, and the fundamental needs and goals that drive their use of data. In light of these commonalities, agency representatives interviewed relied heavily on informal networks of their counterparts in peer agencies to learn about new approaches and troubleshoot issues. Agency staff connected with these networks through professional organizations (e.g., TRB, APTA), research publications, and the use of common vendors.

References

1. B. Hemily, "The Use of Transit ITS Data for Planning and Management, and Its Challenges; a Discussion Paper." ITS America, 2015.
2. S. Descant, "Data, Analysis Are Essential to Planning the Future of Transit," 2021.
3. IBM, "Data Science," 15 May 2020. https://www.ibm.com/cloud/learn/data-science-introduction.
4. U. Fayyad and H. Hamutcu, "Toward Foundations for Data Science and Analytics: A Knowledge Framework for Professional Standards." *Harvard Data Science Review*, vol. 2, no. 2, 2020.
5. National Academies of Sciences, Engineering, and Medicine, *Guidebook for Managing Data from Emerging Technologies for Transportation*. The National Academies Press, Washington, 2020.
6. American Public Transportation Association, *Policy Brief: Leveraging Big Data in the Public Transportation Industry*. American Public Transportation Association, 2019.
7. National Academies of Sciences, Engineering, and Medicine, *Data Sharing Guidance for Public Transit Agencies – Now and in the Future*. The National Academies Press, Washington, 2020.
8. National Academies of Sciences, Engineering, and Medicine, *The Transit Analyst Toolbox: Analysis and Approaches for Reporting, Communicating, and Examining Transit Data*. The National Academies Press, Washington, 2021.
9. Shared-Use Mobility Center, "Objective-Driven Data Sharing for Transit Agencies in Mobility Partnerships." Shared-Use Mobility Center, Chicago, 2019.
10. T. F. Welch and A. Widita, "Big Data in Public Transportation: A Review of Sources and Methods." *Transport Reviews*, vol. 39, no. 6, pp. 795-818, 2019.
11. Organisation for Economic Co-operation and Development, "Glossary of Statistical Terms." 8 March 2006. https://stats.oecd.org/glossary/detail.asp?ID=532.
12. A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers." *IBM Journal of Research and Development*, vol. 3, no. 3, pp. 210-229, July 1959.
13. S. Brown, "Machine Learning, Explained," 21 April 2021. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained.
14. University at Buffalo Center for Multisource Information Fusion, "What is Multisource Information Fusion?" http://www.buffalo.edu/cmif/center/what-is-MIF.html.
15. T. Meng, X. Jing, Z. Yan and W. Pedrycz, "A Survey on Machine Learning for Data Fusion." *Information Fusion*, vol. 57, pp. 115-219, 2020.
16. Cornell University Computer Vision and Image Analysis Lab, "Computer Vision," 2019. http://www.via.cornell.edu/research/overview.html#.
17. D. Gettman, "Raising Awareness of Artificial Intelligence for Transportation Systems." Federal Highway Administration, Washington, 2019.
18. D. Otter, J. R. Medina and J. K. Kalita, "A Survey of the Usages of Deep Learning for Natural Language Processing." *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, pp. 604-624, 2021.
19. S. Rangasamy, R. Nadenichek, M. Rayasam and A. Sozdatelev, "National Language Processing in Healthcare," 6 December 2018. https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/natural-language-processing-in-healthcare.

20. L. Wright and S. Davidson, "How to Tell the Difference Between a Model and a Digital Twin." *Advanced Modeling and Simulation in Engineering Sciences*, vol. 7, no. 13, 2020.
21. A. Parrott, B. Umbendhauer and L. Warshaw, "Digital Twins: Bridging the Physical and Digital," 15 January 2020. https://www2.deloitte.com/us/en/insights/focus/tech-trends/2020/digital-twin-applications-bridging-the-physical-and-digital.html.
22. M. Wanek-Libman, "More than a Mirror Image: Utilizing Digital Twins Can Boost the Management and Monitoring of Rail Assets." *Mass Transit*, 7 February 2019.
23. J. Taylor, P. Culligan, S. Derrible, R. Jain and D. Lam, "Smart City DIgital Twin Convergence Workshop: Final Report." Georgia Institute of Technology, Atlanta, 2019.
24. M. Wanek-Libman, "More than a Mirror Image." *Mass Transit*, 7 February 2019.
25. "Preteckt," 2020. https://www.preteckt.com/.
26. Partnership for New York City and Metropolitan Transportation Authority, "Transit Tech Lab," 2021. https://transitinnovation.org/lab.
27. RFID Ready, "Predictive Maintenance and Automated Asset Management for Transit Authority." https://www.rfidready.net/files/131025862.pdf.
28. Federal Railroad Administration, "Robust Anomaly Detection for Vision-Based Inspection of Railway Components." Washington, 2015.
29. A. Adler, "Future View: United Road Best on Uptake Predictive Maintenance," 20 November 2020. https://www.freightwaves.com/news/seeing-the-future-united-road-bets-on-uptake-predictive-maintenance.
30. S. Brawner, "Fleets Move Toward Predictive Maintenance to Prevent Breakdowns, Reduce Expenses," 9 March 2020. https://www.ttnews.com/articles/fleets-move-toward-predictive-maintenance-prevent-breakdowns-reduce-expenses.
31. J. Melin, "Moving Beyond the Hype of Predictive Maintenance," Honewell Aerospace, 2021. https://aerospace.honeywell.com/us/en/learn/about-us/blogs/moving-beyond-the-hype-of-predictive-maintenance.
32. Rolls Royce, "Predictive Maintenance – Pulse," 2021. https://www.rolls-royce.com/products-and-services/nuclear/nuclear-lifecycle/operations-and-lifetime-extensions/ic-long-term-support/predictive-maintenance-pulse.aspx#/.
33. R. Johnston, "Miami-Dade Transit Using AI to Monitor Passengers for Social Distancing," 10 November 2020. https://statescoop.com/miami-dade-transit-using-ai-to-monitor-passengers-for-social-distancing/.
34. R. Johnston, "Miami-Dade Transit Using AI to Monitor Passengers for Social Distancing." *StateScoop*, 10 November 2020.
35. M. Luczak, "LIRR Travel App Now Offers "Crowding Data." *Railway Age*, 14 September 2020.
36. J. D. Antos, W. Jia and J. H. Parker, "Is It Too Crowded in Here? In Search of Safety Standards for Pedestrian Congestion in Rail Stations?" *Transportation Research Record*, vol. 2648, no. 1, pp. 126-133, 2017.
37. B. Wai and W. Zhou, "Designing and Implementing Real-Time Bus Time Predictions Using Artificial Intelligence." *Transportation Research Record*, vol. 2674, no. 11, pp. 636-648, 2020.

38. A. Roman, "Innovation Solutions: Operations," 19 October 2020.
39. National Center for Transit Research, "Transit Extraboard Management – Optimum Sizing & Strategies." University of South Florida, Center for Transportation Research, Tampa, 2007.
40. J. G. Strathman, J. Broach and S. Callas, "Evaluation of Short Duration Unscheduled Absences Among Transit Operators: TriMet Case Study." Transportation Research and Education Center, Portland, 2009.
41. Verint, "Helping a Railroad Service Conduct Business." https://www.verint.com/wp-content/uploads/CS_Amtrak_US_Final_0720.pdf.
42. U.S. Customs and Border Protection, "Biometrics." https://biometrics.cbp.gov/.
43. "Progress Rail Fatigue Monitoring Enhances Operator Safety." *Progressive Railroading*, 21 August 2017.
44. H. Golub, "StreetLight Releases On-Demand Bus and Rail Metrics for Transit," 4 February 2021. https://www.streetlightdata.com/bus-and-rail-metrics/?type=blog/.
45. W. Qin, Y.-N. Sun, Z.-L. Zhuang, Z.-Y. Lu and Y.-M. Zhou, "Multi-agent Reinforcement Learning-based Dynamic Task Assignment for Vehicles in Urban Transportation System." *International Journal of Production Economics*, vol. 240, p. 108251, 2021.
46. D. Luo, D. Zhao, Q. Ke, X. You, L. Liu, D. Zhang, H. Ma and X. Zuo, "Fine-Grained Service-Level Passenger Flow Prediction for Bus Transit Systems Based on Multitask Deep Learning." *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 11, pp. 7184-7199, 2021.
47. T. Tang, A. Fonzone, R. Liu and C. Choudhury, "Multi-stage Deep Learning Approaches to Predict Boarding Behaviour of Bus Passengers." *Sustainable Cities and Society*, vol. 73, 2021.
48. P. Noursalehi, H. N. Koutsopoulos and J. Zhao, "Dynamic Origin-Destination Prediction in Urban Rail Systems: A Multi-Resolution Spatio-Temporal Deep Learning Approach." *IEEE Transactions on Intelligent Transportation Systems*, pp. 1-10, 2021.
49. P. Noursalehi, H. N. Koutsopoulos and J. Zhao, "Machine-Learning-Augmented Analysis of Textual Data: Application in Transit Disruption Management." *IEEE Open Journal of Intelligent Transportation Systems*, vol. 1, pp. 227-236, 2020.
50. S. Arhin, B. Manandhar, H. Baba Adam and A. Gatiba, "Predicting Bus Travel Times in Washington, DC Using Artificial Neural Networks (ANNs)." Mineta Transportation Institute, San Jose, 2021.
51. TransitCenter, "The Data Transit Riders Want: A Shared Agenda for Public Agencies and Transit Application Developers." TransitCenter, New York, 2018.
52. "TRB Standing Committee on Transit Data – AP090." http://www.trb-transit-data.org/.
53. Transportation Research Board, "TCRP G-19: Improving Access and Management of Transit ITS Data." https://apps.trb.org/cmsfeed/TRBNetProjectDisplay.asp?ProjectID=4687.
54. National Institute of Standards and Technology, "NIST Big Data Interoperability Framework: Volume 4, Security and Privacy." National Institute of Standards and Technology, Washington, 2019.
55. P. K. Rajesh, N. Manikandan, C. S. Ramshankar, T. Vishwanathan and C. Sathishkumar, "Digital Twin of an Automotive Brake Pad for Predictive Maintenance." *2nd International Conference on Recent Trends in Advanced Computing*, Chennai, 2020.

Acronyms and Abbreviations

AFC	Automated Fare Collection
AI	Artificial Intelligence
AIM	Accelerating Innovative Mobility
APC	Automated Passenger Counter
APTA	American Public Transportation Association
AVL	Automatic Vehicle Location
BI	Business Intelligence
CAN	Controller Area Network
FRA	Federal Railroad Administration
FTA	Federal Transit Administration
GPS	Global Positioning System
GTFS	General Transit Feed Specification
IMI	Integrated Mobility Integration
IT	Information Technology
ITS	Intelligent Transportation Systems
LBS	Location-Based Services
ML	Machine Learning
MOD	Mobility on Demand
NLP	Natural Language Processing
RFID	Radio-Frequency Identification
TCRP	Transit Cooperative Research Program
TRB	Transportation Research Board
TRID	Transport Research International Documentation
USDOT	United States Department of Transportation

U.S. Department of Transportation

Federal Transit Administration

U.S. Department of Transportation
Federal Transit Administration
East Building
1200 New Jersey Avenue, SE
Washington, DC 20590
https://www.transit.dot.gov/about/research-innovation